Breast Cancer Classification

Salih Eren Yüzbaşıoğlu / AI AR-GE Club / 19.02.2024

TABLE OF CONTENTS

Introduction

In this assignment, our objective is to conduct a thorough investigation of a comprehensive dataset comprising valuable information derived from Dr. Wolberg's extensive collection of clinical case reports. The dataset encapsulates a wide array of essential attributes and variables pertaining to individuals, meticulously gathered and analyzed over a significant period. Our primary focus is to meticulously examine and discern patterns, trends, and correlations within the dataset, particularly regarding the presence or absence of malignancy in cancer cases.

Exploratory Data Analysis

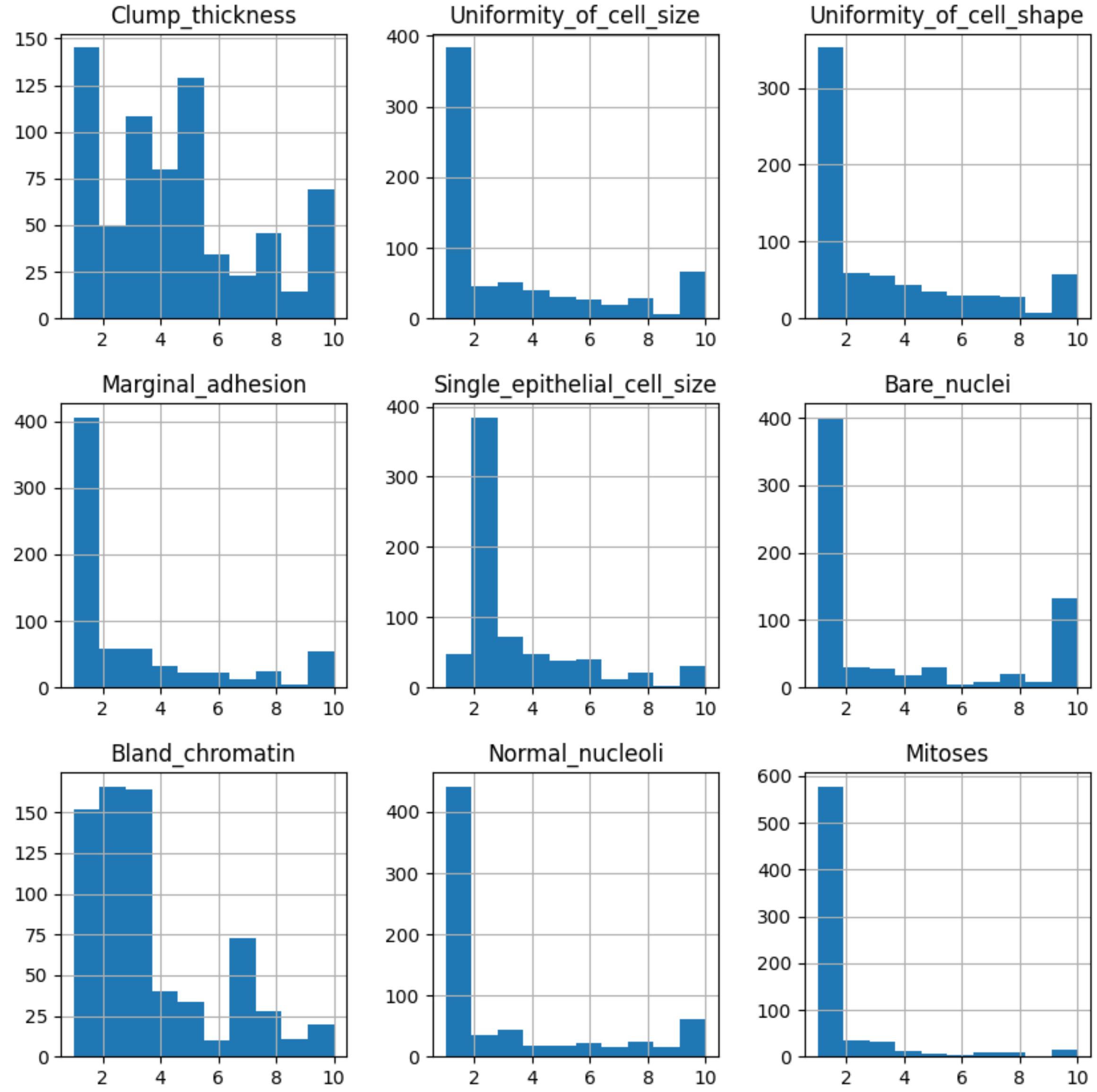
The data set under investigation was meticulously collected by Dr. Wolberg during the period spanning 1989 to 1991 and comprises a total of 699 distinct clinical cases. Upon initial examination of the data, two noteworthy observations came to light. Firstly, the data is not provided in a metric format, as each column is labeled on a scale from 1 to 10, yet no explicit information is available regarding the specific criteria used to assign these numerical values. Secondly, the dataset itself contains only 699 entries, which may present limitations when attempting to construct robust predictive models for determining the malignancy of future cancer cases. The relatively small number of data points across the eight different feature columns may impact the strength and reliability of such predictive models.

Additionally, another aspect that caught my attention is the absence of details regarding the measurement methods employed to capture the various features. Considering that the data collection took place in 1990, there is a concern regarding the precision and accuracy of the measurements obtained. The lack of information regarding the specific techniques utilized raises potential questions regarding the reliability and consistency of the recorded features.

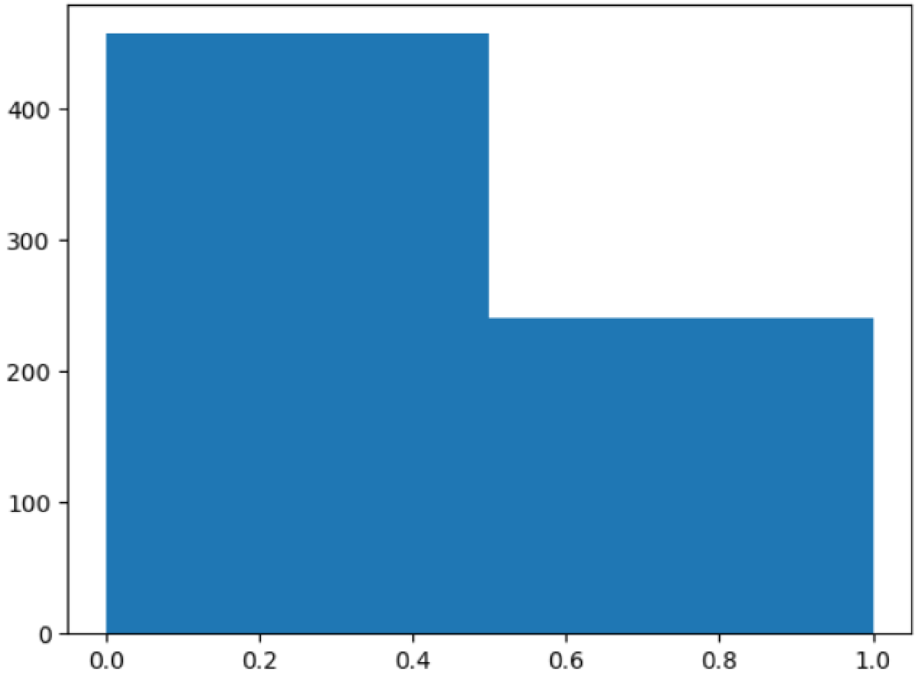
Further investigation of the data set's website reveals that it has been referenced in six distinct research papers. However, upon closer examination, it becomes evident that the primary utilization of the dataset was not for training models to predict cancer malignancy in future cases. Instead, it has primarily served as a resource for testing and evaluating various statistical analysis methods.

These insights shed light on the dataset's characteristics and its potential limitations. While it presents an invaluable resource for exploring statistical techniques, caution must be exercised when extrapolating the findings to real-world predictive applications. It is crucial to acknowledge the dataset's specific purpose and ensure that any conclusions drawn align with its intended usage in statistical analysis rather than direct clinical prediction.

Below figure has histogram for every feature in data set.



Upon a visual examination of the histograms, a notable observation arises. It becomes apparent that certain columns, namely Normal_nucleoli, Mitoses, Bare_nuclei, Single…, Marginal_adhesion, Uniformity_of_cell_size, and Uniformity_of_cell_shape, exhibit a predominant occurrence of a single value throughout the dataset. This particular trend poses challenges for our exploratory data analysis (EDA) efforts, as rows sharing similar features do not provide substantial insights.



Looking at the above graph we can see that there are more non-malignant data than there are malignant which is good as while predicting a cell to be malignant is important not wrongfully predicting is as important as well. Wide range of non-malignant data will help our machine learning models to train better.

Deleting Null Values

Looking at the data set's website we see that it is given to use which columns have null values. It is only Bare Nuclei column that have null values. Quick scan reveals there are only 16 null values which is not a whole lot for 699 values. Looking at the histogram of Bare Nuclei we realize it has weird shape values 1 and 10 mostly being 1 capture most of the possible values the column takes. So, it would not make much sense to use mean value to fill null values as mean is not used a whole lot instead, we use mod value and fill null rows with 1 which is the mod.

Accuracy, Precision, Recall and ROC AUC Score

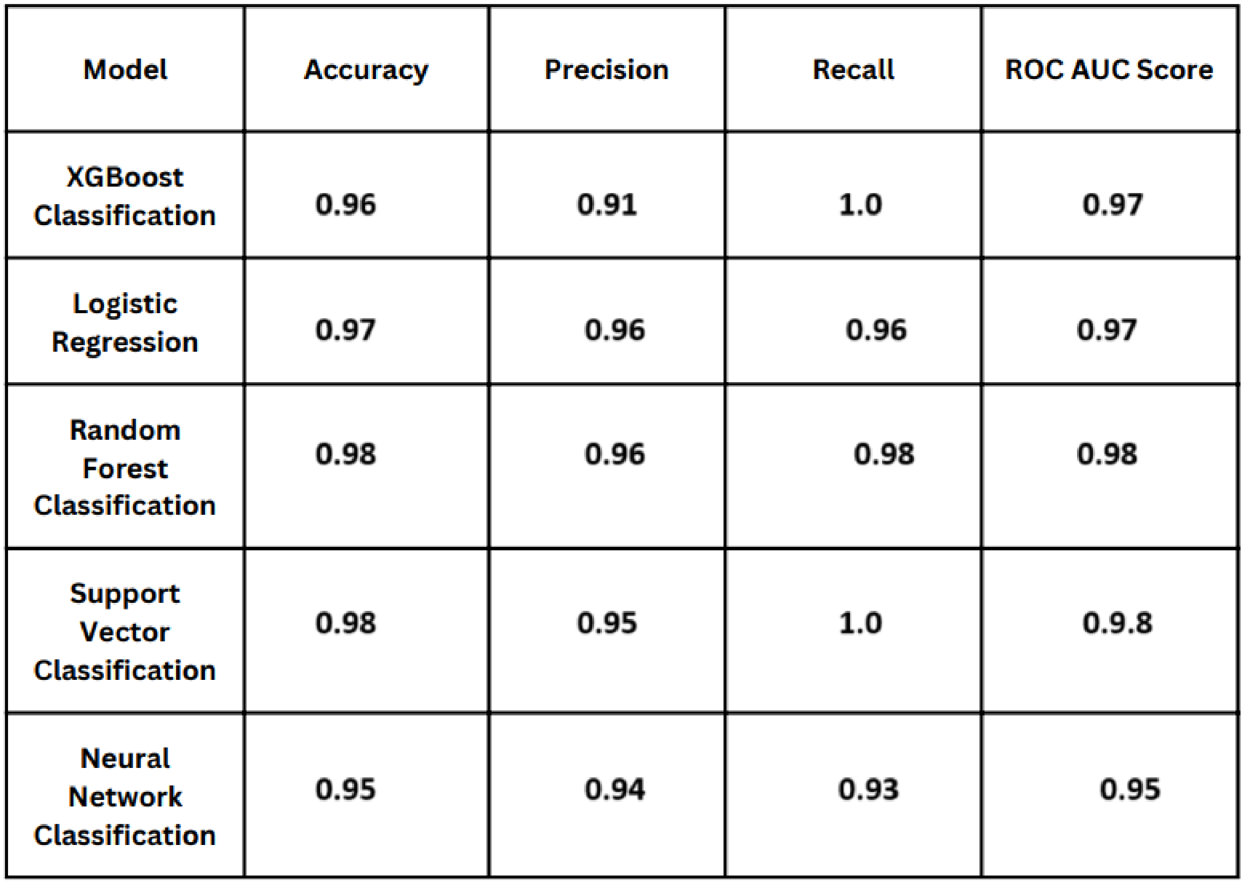
Accuracy: Accuracy is a measure of how often the model makes correct predictions. It is calculated by dividing the number of correct predictions by the total number of predictions. While accuracy is a straightforward metric, it can be misleading when classes are imbalanced. For example, if 95% of the samples belong to class A and only 5% belong to class B, a naive model that always predicts class A would achieve 95% accuracy, even though it fails to detect any samples from class B.

Precision: Precision is the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive (true positives + false positives). It measures the model's ability to avoid falsely labeling negative samples as positive. Precision is especially important when the cost of false positives is high. For instance, in a medical diagnosis scenario, precision indicates the percentage of correctly identified positive cases out of all the predicted positive cases.

Recall (also known as Sensitivity or True Positive Rate): Recall is the proportion of correctly predicted positive instances (true positives) out of all actual positive instances (true positives + false negatives). It quantifies the model's ability to identify all positive samples. Recall is crucial when the cost of false negatives is high. For example, in a spam email classification task, recall indicates the percentage of correctly detected spam emails out of all the actual spam emails.

ROC AUC Score: ROC (Receiver Operating Characteristic) is a graphical plot that illustrates the performance of a binary classification model at various classification thresholds. The ROC curve is created by plotting the true positive rate (sensitivity/recall) against the false positive rate (1 - specificity). The Area Under the ROC Curve (ROC AUC) summarizes the overall performance of the model across all classification thresholds. A higher ROC AUC score indicates better discrimination and a better ability to distinguish between positive and negative samples.

Model Comparison



Model	Accuracy	Precision	Recall	ROC AUC Score
XGBoost Classification	0.96	0.91	1.0	0.97
Logistic Regression	0.97	0.96	0.96	0.97
Random Forest Classification	0.98	0.96	0.98	0.98
Support Vector Classification	0.98	0.95	1.0	0.9.8
Neural Network Classification	0.95	0.94	0.93	0.95

Looking first time at the table we see every model has a very high score possibly meaning we have some overfit,

Which is very possible as our data only contained 699 rows and lots of columns had a value that was used so much more than other columns.

Before comparing models, we should think which score is the most important, first thing that comes to mind is finding if cancer is malignant is extremely important so we need true positives and we need to minimize false negatives, this is much important than categorizing a non-malignant cancer malignant as extra treatment would not be deadly. So, the score that we are looking for is Recall.

If we look at the models and compare them, we realize neural network performed the worst, achieving lowest score in every domain so we can eliminate it first. Other than that, looking at the table we realize Random Forest beats logistic regression in every domain so we can eliminate logistic regression as well. And similarly, SVC beats XGBoost in every domain as well.

So, we got 2 models left SVC and Random Forest. Between these 2 models there is a tradeoff SVC scores higher in recall while Random Forest scores higher in precision as we mentioned before recall is more important, so we choose SVC as our winner here.

Extra Model

As we talked before lots of columns have values that occur multiple times so, there is a very high chance that there are rows in train data set that very close to a row in test data set in terms of features, which makes KNN a good choice to be extra model.

Here are KNN results:

0.98, 0.96, 0.98, 0.98

Which are basically the same as Random Forest Classification. Which makes KNN a very good alternative to random forest as it trained faster than it as well. But if we are looking for an alternative to SVC, XGBoost looks like a good idea even though it has lower scores the score that we found important, recall, is high when using XGBoost.